Published in: Pre-Proceedings of the International Conference on Linguistic Evidence. Empirical, Theoretical and Computational Perspectives. - Tübingen: University of Tübingen, 2006. Pp. 195-197.

Using Corpus Statistics in the Modeling of Linguistic Paradigms

Beata Trawiński

SFB 441 – University of Tübingen, Germany

trawinski@sfs.uni-tuebingen.de

This paper presents how corpus statistics can be used to verify complex inflectional paradigms. This will be demonstrated using a set of traditionally assumed inflectional paradigms of third person personal pronouns in Polish.

According to traditional approaches, the inventory of third person personal pronouns in Polish comprises entities specified in (1)–(5). Four grammatical categories play a role in constituting the inflectional paradigm of the pronouns, i.e., case (nominative, genitive, dative, accusative, instrumental and locative), number (singular and plural), postprepositionality (yes or no; see forms separated by a slash in (1)–(5)) and accentability (yes or no; indicated by pronoun forms in parentheses versus forms not in parentheses in (1)–(5)). In (1)–(5), only genitive and accusative masculine singular pronouns are specified as being able to appear in the unaccented postprepositional form (cf. the form *-ń* in (1), (2) and (3)). However, the corpus data seem to demonstrate more possibilities (for our investigations, the IPI PAN Corpus of Polish was used, cf. http://korpus.pl).

Table 1 presents the distribution of selected unaccented postprepositional pronouns found in the corpus. For each form, the context in which it occurs is specified, i.e., the contraction of that form with a particular preposition, and the total number of times it occurred together with the percentage of the total frequency of all unaccented postprepositional forms is recorded. In addition, the total of all occurrences of each contraction found in the corpus is indicated, as well as the percentage of the total frequency of all preposition-pronoun contractions occurring in the corpus is calculated. To determine whether the distribution of the unaccented postprepositional pronouns in the corpus may be considered linguistically significant and, in consequence, may establish a basis for the revision of the inflectional paradigms in (1)–(5), a number of quantitative procedures has been performed. In our ongoing study, the distribution of genitive and accusative feminine singular, as well as plural postprepositional pronouns has been analyzed. In future work, frequencies of the remaining pronouns will be examined, and, based on the evaluation results, the correctness of the inflectional paradigms in (1)–(5) will be confirmed or challenged.

(1) Inflectional paradigm of the third person masculine human pronoun *on* ‘he’:

CASE	SINGULAR	PLURAL
nominative	*on* (—) / — (—)	*oni* (—) / — (—)
genitive	*jego (go)* / *niego (-ń)*	*ich* (—) / *nich* (—)
dative	*jemu (mu)* / *niemu* (—)	*im* (—) / *nim* (—)
accusative	*jego (go)* / *niego (-ń)*	*ich* (—) / *nich* (—)
instrumental	*nim* (—) / *nim* (—)	*nimi* (—) / *nimi* (—)
locative	— (—) / *nim* (—)	— (—) / *nich* (—)

(2) Inflectional paradigm of the third person masculine animal pronoun *on* ‘it’:

CASE	SINGULAR	PLURAL
nominative	*on* (—) / — (—)	*one* (—) / — (—)
genitive	*jego (go)* / *niego (-ń)*	*ich* (—) / *nich* (—)
dative	*jemu (mu)* / *niemu* (—)	*im* (—) / *nim* (—)
accusative	*jego (go)* / *niego (-ń)*	*je* (—) / *nie* (—)
instrumental	*nim* (—) / *nim* (—)	*nimi* (—) / *nimi* (—)
locative	— (—) / *nim* (—)	— (—) /*nich* (—)

(3) Inflectional paradigm of the third person masculine inanimate pronoun *on* ‘it’:

CASE	SINGULAR	PLURAL
nominative	*on* (—) / — (—)	*one* (—) / — (—)
genitive	*jego (go)* / *niego (-ń)*	*ich* (—) / *nich* (—)
dative	*jemu (mu)* / *niemu* (—)	*im* (—) / *nim* (—)
accusative	*jego (go)* / *niego (-ń)*	*je* (—) / *nie* (—)
instrumental	*nim* (—) / *nim* (—)	*nimi* (—) / *nimi* (—)
locative	— (—) / *nim* (—)	— (—) / *nich* (—)

(4) Inflectional paradigm of the third person feminine pronoun *ona* ‘she’:

CASE	SINGULAR	PLURAL
nominative	*ona* (—) / — (—)	*one* (—) / — (—)
genitive	*jej* (—) / *niej* (—)	*ich* (—) / *nich* (—)
dative	*jej* (—) / *niej* (—)	*im* (—) / *nim* (—)
accusative	*ją* (—) / *nią* (—)	*je* (—) / *nie* (—)
instrumental	*nią* (—) / *nią* (—)	*nimi* (—) / *nimi* (—)
locative	— (—) / *niej* (—)	— (—) / *nich* (—)

(5) Inflectional paradigm of the third person neuter pronoun *ono* ‘it’:

CASE	SINGULAR	PLURAL
nominative	*ono* (—) / — (—)	*one* (—) / — (—)
genitive	*jego (go)* / *niego* (—)	*ich* (—) / *nich* (—)
dative	*jemu (mu)* / *niemu* (—)	*im* (—) / *nim* (–)
accusative	*je* (—) / *nie* (—)	*je* (—) / *nie* (—)
instrumental	*nim* (—) / *nim* (—)	*nimi* (—) / *nimi* (—)
locative	— (—) / *nim* (—)	— (—) / *nich* (—)

	dlań 'for.PR'	doń 'to.PR'	nań 'on.PR'	weń 'in.PR'	zeń 'with.PR'	odeń 'from.PR'	przezeń 'by.PR'	poń 'after.PR'	zań 'behind.PR'	przedeń 'in front of.PR	Total, %
[...]											[...] [...]
nom, n, sg											0 0.00 %
gen, n, sg	3	16			16	1					36 3.02 %
dat, n, sg											0 0.00 %
acc, n, sg			13	6			32		2		53 4.45 %
instr, n, sg											0 0.00 %
loc, n, sg											0 0.00 %
nom, n, pl											0 0.00 %
gen, n, pl		5									5 0.42 %
dat, n, pl											0 0.00 %
acc, n, pl				1			1				2 0.17 %
instr, n, pl											0 0.00 %
loc, n, pl				1							1 0.08 %
nom, f, sg											0 0.00 %
gen, f, sg	5	15			4	1					25 2.06 %
dat, f, sg											0 0.00 %
acc, f, sg			5	4			10				19 1.59 %
instr, f, sg											0 0.00 %
loc, f, sg											0 0.00 %
nom, f, pl											0 0.00 %
gen, f, pl	1	1			2	1					5 0.42 %
dat, f, pl											0 0.00 %
acc, f, pl			2				1				3 0.25 %
instr, f, pl											0 0.00 %
loc, f, pl			1								1 0.08 %
Total	101	219	377	101	93	23	250	1	27	1	1.193
%	8.47%	18.36%	31.60%	8.47%	7.80%	1.93%	20.96%	0.08%	2.26%	0.08%	100%

Table 1: The distribution of selected postprepositional unaccented pronouns in the IPI PAN Corpus